Scientific Research Publishing

Open Journal of Statistics, 2017, 7, 761-767
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Estimating the Empirical Null Distribution of Maxmean Statistics in Gene Set Analysis

Xing Ren[1], **Jianmin Wang**[2], **Song Liu**[2], **Jeffrey C. Miecznikowski**[1*]

[1]Department of Biostatistics, University at Buffalo, Buffalo, USA
[2]Department of Biostatistics and Bioinformatics, Roswell Park Cancer Institute, Buffalo, USA
Email: xingren@buffalo.edu, Jianmin.Wang@roswellpark.org, song.liu@roswellpark.org, *jcm38@buffalo.edu

How to cite this paper: Ren, X., Wang, J.M., Liu, S. and Miecznikowski, J.C. (2017) Estimating the Empirical Null Distribution of Maxmean Statistics in Gene Set Analysis. *Open Journal of Statistics*, **7**, 761-767.
https://doi.org/10.4236/ojs.2017.75053

Received: May 22, 2017
Accepted: September 22, 2017
Published: September 27, 2017

Abstract

Gene Set Analysis (GSA) is a framework for testing the association of a set of genes and the outcome, e.g. disease status or treatment group. The method replies on computing a maxmean statistic and estimating the null distribution of the maxmean statistics via a restandardization procedure. In practice, the pre-determined gene sets have stronger intra-correlation than genes across sets. This may result in biases in the estimated null distribution. We derive an asymptotic null distribution of the maxmean statistics based on sparsity assumption. We propose a flexible two group mixture model for the maxmean statistics. The mixture model allows us to estimate the null parameters empirically via maximum likelihood approach. Our empirical method is compared with the restandardization procedure of GSA in simulations. We show that our method is more accurate in null density estimation when the genes are strongly correlated within gene sets.

Keywords

Gene Set Analysis, Maxmean, Empirical Null, Mixture Model

1. Introduction

A gene pathway commonly refers to a set of genes that share a particular property, carry out a biological function or lead to a certain product in cells/tissues. Performing differential expression (DE) analysis on such gene sets aggregates the signal of individual genes and potentially increases the power of a hypothesis test. Gene set analysis also provides comprehensive understanding of the biological activities associated with the outcome phenotype and may shed light on treatment of disease.

A variety of tools are available for gene set analysis. These methods can be roughly classified into two broad categories, self-contained and competitive [1]. The self-contained methods test the association between the phenotype and the gene set while ignoring the other genes. Competitive methods overcome this limitation by taking into account other genes when evaluating the association. A common approach is to compute the gene level statistics, and then aggregate them into a gene-set level summary statistic. Of the many competitive methods gene set enrichment analysis (GSEA) [2] and (GSA) [3] are two representative algorithms. In GSEA the Kolmogorov-Smirnov (KS) statistic is employed. GSA improves the power of GSEA by using a more powerful maxmean statistic. First the gene level z statistics are calculated, $z_i, i = 1, \cdots, n$. Let $\mathcal{S}$ denote the indices of the gene set and $n_\mathcal{S}$ be the size of $\mathcal{S}$. The maxmean statistic S is defined as,

$$\begin{aligned} S_+ &= \frac{1}{n_\mathcal{S}} \sum_{i \in \mathcal{S}} z_i I\{z_i > 0\}, \\ S_- &= -\frac{1}{n_\mathcal{S}} \sum_{i \in \mathcal{S}} z_i I\{z_i < 0\}, \\ S &= \max(S_+, S_-), \end{aligned} \tag{1.1}$$

where $I\{\}$ is the indicator function.

The GSA method estimates the null distribution of S through a restandardization procedure, which is a combination of row (gene) randomization and column (sample label) permutation. It compares the gene set against its permutations and also takes into account the overall distribution of randomly selected null sets. The permutation maintains the correlation structure in the gene set, while row randomization rescales and shifts the permutations to include the competition of genes outside the set.

In practice, the gene sets for analysis are obtained from a public database e.g. Kyoto Encyclopedia of Genes and Genomes (KEGG) [4], MSigDB [2] or gene ontology (GO) [5]. It is reasonable to expect that genes in the same set have stronger correlation than genes across different sets. In such cases there may be a mismatch from the null distribution estimated from restandardization and the true null, which leads to biased inference.

We propose a two group mixture model for the gene set maxmean statistics. Our model assumes that only a small proportion of the gene sets are truly significantly DE. Based on the mixture model, we apply the maximum likelihood method to estimate the empirical null. The empirical method improves the accuracy of GSA in large scale hypothesis testing. It also reduces the computational burden of the permutation steps in restandardization procedure of GSA. The analysis is demonstrated in simulation studies and a data set from the MSigDB database [2].

2. Methods

The maxmean statistic in GSA is essentially a maximum statistic of two correlated sample means, S_+ and S_- defined in (1.1), which asymptotically follow bivariate normal distribution for adequately large $n_\mathcal{S}$,

$$\begin{pmatrix} S_+ \\ S_- \end{pmatrix} \sim N_2 \left\{ \begin{pmatrix} \mu_+ \\ \mu_- \end{pmatrix}, \begin{pmatrix} \sigma_+^2 & \rho\sigma_+\sigma_- \\ \rho\sigma_+\sigma_- & \sigma_-^2 \end{pmatrix} \right\}, \tag{2.1}$$

where $\mu_+ = \mathrm{E}(S_+)$, $\sigma_+^2 = \mathrm{Var}(S_+)$, $\mu_- = \mathrm{E}(S_-)$, $\sigma_-^2 = \mathrm{Var}(S_-)$, and $\rho = \mathrm{corr}(S_+, S_-)$. Based on the work in [6] [7], an asymptotic distribution of the maxmean statistics under the Lindeberg condition (see [8]) is

$$\begin{aligned} f_0(s) = {} & \frac{1}{\sigma_+}\phi\left(\frac{-s+\mu_+}{\sigma_+}\right) \times \Phi\left(\frac{\rho(-s+\mu_+)}{\sigma_+\sqrt{1-\rho^2}} - \frac{-s+\mu_-}{\sigma_-\sqrt{1-\rho^2}}\right) \\ & + \frac{1}{\sigma_-}\phi\left(\frac{-s+\mu_-}{\sigma_-}\right) \times \Phi\left(\frac{\rho(-s+\mu_-)}{\sigma_-\sqrt{1-\rho^2}} - \frac{-s+\mu_+}{\sigma_+\sqrt{1-\rho^2}}\right) \end{aligned} \tag{2.2}$$

where ϕ and Φ are the probability density function (pdf) and cumulative density function (cdf) of the standard normal distribution.

We can estimate the parameters in f_0 by fitting the null gene sets. A special case would be that $z_i \sim N(0, \sigma^2)$ independently. We can easily compute the parameters:

$$\mu_+ = \mu_- = 0.40\sigma, \sigma_+^2 = \sigma_-^2 = 0.34\sigma^2, \rho = -0.467.$$

However, genes in the same set are often correlated. Therefore, the theoretically computed parameters may not match the actual null distribution well. We propose an empirical method to estimate the null distribution of *S*. Let f be the density function of the maxmean statistics for all the gene sets. Adopting the two group mixture model [9] [10], we specify a similar model that f is comprised by a large proportion (p_0) of the null density f_0 and a small proportion of the non-null density f_1,

$$f(s) = p_0 f_0(s) + p_1 f_1(s). \tag{2.3}$$

The null density f_0 is assumed to have the form in (2.2), in which the parameters (μ_+, μ_-, σ_+, σ_-, ρ) need to be estimated. For identifiability of f_0 we further assume the non-null density $f_1(s) \approx 0\ \forall s \in A_0$ for some interval A_0. Under this assumption, *S* follows a truncated distribution $f_T(s)$ for $s \in A_0$,

$$f_T(s) \approx \frac{p_0 f_0(s)}{\int_{A_0} p_0 f_0(s)} = \frac{f_0(s)}{\int_{A_0} f_0(s)}. \tag{2.4}$$

Fitting the maxmean statistics in A_0 to (2.4) by maximum likelihood yields $\hat{f}_0$. In addition, p_0 can be estimated by

$$\hat{p}_0 = \frac{\#\{s_i \in A_0\}}{n\int_{A_0} \hat{f}_0(s)}. \tag{2.5}$$

In this paper we let A_0 be the interval (0, q_S) where q_S is 90% quantile of the maxmean statistics of all gene sets.

3. Simulations

We simulate 2000 gene sets, 5% of them are DE sets and the other 95% are null sets. All sets contain $n_S = 50$ genes. Let C_1 and C_2 be the sample indices of two conditions and each condition has m = 10 samples. For gene i in sample j, the expression data x_{ij} is generated in a hierarchical fashion,

$$x_{0kj} \sim N\left(\delta_k I\left\{j \in C_1\right\}, \tau_k^2\right),$$
$$x_{ij} \sim N\left(\alpha_i x_{0kj}, \sigma_i^2\right) \forall i \in \mathcal{S}_k,$$

where $\tau_k = \sigma_i = 1$, $\delta_k = 1$ for DE sets and $\delta_k = 0$ for null nets. x_{0kj} is the expression of the hub gene [11] of set S_k and all genes in the same set are correlated with the hub gene. The DE genes of set S_k are jointly controlled by δ_k and α_i. In particular, δ_k controls the differential expression of the hub gene between two conditions. The parameter α_i controls the inter-correlation within the set. When $\alpha_i = 0$ genes are independent within gene sets. For null sets we let $\alpha_i = 0$ (independent) or ±0.2 (correlated). For DE sets $\alpha_i = \pm 1$ so that the correlation is stronger than the null sets. **Figure 1** shows the f_0 density estimate by our empirical method and the restandardization procedure in GSA.

When genes within the gene sets are independent, the null distribution obtained by the two methods are very close (**Figure 1** left). As 95% of the gene sets are null sets, the estimated distributions match the majority of the sets. When genes are correlated, the null distribution obtained by GSA shows a mismatch from the null gene sets, while our empirical method maintains a good fit to the data (**Figure 1** right).

4. Application

We evaluate our empirical method and GSA with the 4722 curated gene sets in MSigDB database using the gender dataset included in the GSEA package [2]. The minimum size of the gene sets is 5, the maximum size is 1972 and median size is 39. The dataset contains transcriptional profiles of 15,056 genes in 17 female and 15 male lymphoblastoid cell line samples. There are 302 gene sets with unadjusted p-value < 0.05 by our empirical method and 507 gene sets by GSA. Controlling the false discovery rate (FDR) at 0.1 with Benjamini-Hochberg pro-

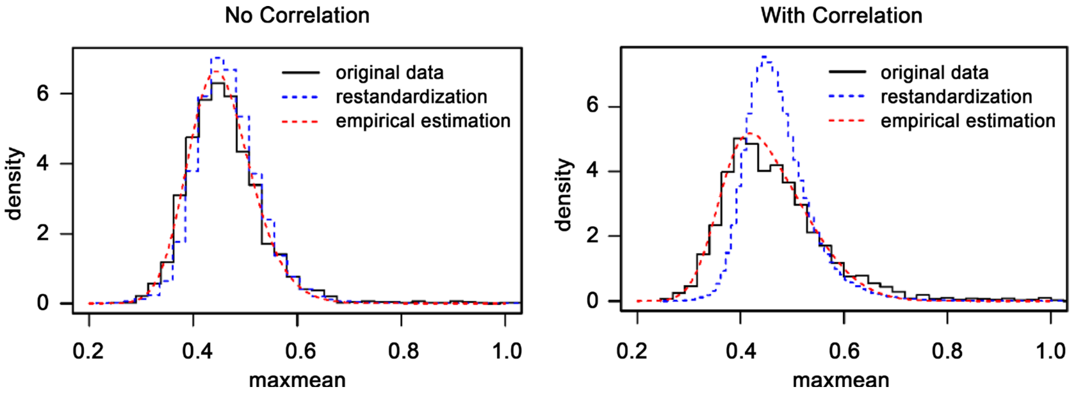

Figure 1. Density estimate by restandardization and empirical method. For null sets, $\delta_k = 0$, $\alpha_i = 0$ (left) and $\alpha_i = \pm 0.2$ (right). For DE sets, $\delta_k = 1$ and $\alpha_i = \pm 1$. The null proportion $\hat{p}_0$ is 0.97 (left) and 0.98 (right) estimated by the empirical method.

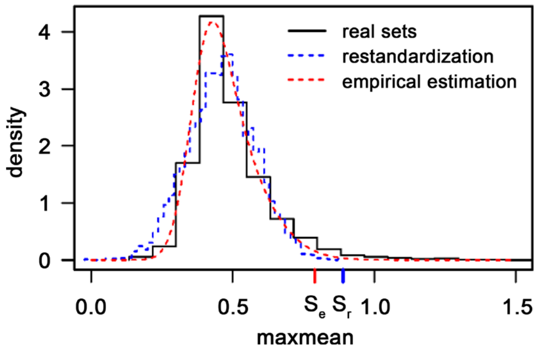

Figure 2. Estimated null distribution of maxmean statistics of the 4722 GSEA gene sets on the gender dataset. $\hat{p}_0 = 0.96$ estimated by the empirical method. The point s_e (red) and s_r (blue) are the cutoff values for the empirical and restandardization methods, respectively at FDR = 0.10.

cedure [12], the empirical method identifies 39 significant sets while the restandardization identifies 8 significant sets. **Figure 2** shows the null distribution obtained by the empirical method and GSA.

5. Discussion

GSA is a representative tool for gene set DE analysis. Compared with the state-of-the-art method GSEA, the maxmean statistic used in GSA is more powerful than the KS statistic in GSEA [3]. GSA also establishes a restandardization algorithm to assess the maxmean statistic. Due to the correlation of the genes in pre-defined gene sets or pathways, the null distribution obtained by the restandardization procedure in GSA may not match the true null distribution well.

In studying the GSA method, we propose a new method to estimate the null distribution of the gene set maxmean statistics. Unlike the permutation test of GSA in which every gene set is compared against its own permutations, the fundamental difference of our method is that it estimates an overall null distribution f_0 parametrically and all gene sets are compared against f_0. The possibility of parametric estimation of f_0 is rendered by the large number of gene sets in the hypothesis testing. A similar idea is proposed in [9] for large-scale test of DE genes. We extend the idea to the test of gene sets when a large number of sets are available.

Our method is based on the sparsity assumption that only a small proportion of the gene sets are truly DE. Further, we adopt a two group parametric mixture distribution to model the gene set maxmean statistics. The parameters of the null distribution is estimated under the sparsity assumption. We show that our new method provides more accurate estimation of the null distribution. It also avoids the computational intensity in the permutation steps of GSA.

In the simulations, we compare the two methods under independence and correlation. When the intra-set correlation is greater than the cross-set correlation, the GSA shows a mismatch to the true null. The reason is that the randomization step samples genes in the entire dataset, which has a different correlation structure from gene sets. As a result the mean and standard deviation obtained in the randomization step is inaccurate.

In application to the gender data set, our method has fewer gene sets with unadjusted p-value < 0.05, but identifies more gene sets after controlling for FDR. A reason is that in GSA, gene-set p-values are limited by the number of permutations. Due to the large number of genes and gene sets, with 10,000 permutations, GSA takes 64 minutes on an Intel i7 3770K 3.5 GHz CPU. In contrast, our empirical method takes less than 1 minute.

An important aspect of our empirical estimation method is specifying an appropriate range for the zero assumption region A_0. In this paper we arbitrarily choose A_0 to be the interval $(0, q_S)$, where q_S is some quantile of the observed maxmean statistics. Determination of q_S is a trade-off between variance and bias. With small q_S, the bias of the null estimate is small but the variance is large and vice-versa. How to determine an optimal A_0 interval requires further exploration.

References

[1] Nam, D. and Kim, S.-Y. (2008) Gene-Set Approach for Expression Pattern Analysis. *Briefings in Bioinformatics*, **9**, 189-197. https://doi.org/10.1093/bib/bbn001

[2] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.* (2005) Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550. https://doi.org/10.1073/pnas.0506580102

[3] Efron, B. and Tibshirani, R. (2007) On Testing the Significance of Sets of Genes. *The Annals of Applied Statistics*, **1**, 107-129. https://doi.org/10.1214/07-AOAS101

[4] Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27-30. https://doi.org/10.1093/nar/28.1.27

[5] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000) Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, **25**, 25-29. https://doi.org/10.1038/75556

[6] Basu, A. and Ghosh, J. (1978) Identifiability of the Multinormal and Other Distributions under Competing Risks Model. *Journal of Multivariate Analysis*, **8**, 413-429. https://doi.org/10.1016/0047-259X(78)90064-7

[7] Cain, M. (1994) The Moment-Generating Function of the Minimum of Bivariate Normal Random Variables. *The American Statistician*, **48**, 124-125.

[8] Billingsley, P. (1995) Probability and Measure. 3rd Edition, Wiley Series in Probability and Mathematical Statistics.

[9] Efron, B. (2004) Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, **99**, 96-104. https://doi.org/10.1198/016214504000000089

[10] Efron, B. (2012) Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Volume 1. Cambridge University Press.

[11] Langfelder, P. and Horvath, S. (2008) WGCNA: An R Package for Weighted Corre-

lation Network Analysis. *BMC Bioinformatics*, **9**, 1-13.
https://doi.org/10.1186/1471-2105-9-559

[12] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* (*Methodological*), **57**, 289-300.